**RAPPORT DE PROJET**

**CO2 PREDICT**

**Émissions de polluants, CO2 et caractéristiques des véhicules commercialisés en France en 2013**



Camille Millon

Gilles Ngamenye

Christophe Seuret

Mentor: Dan Cohen

Plan du rapport

[**PRÉSENTATION DU PROJET**](#_5l02ge83ekb9)

[1. Description](#_aztvayirpkoj)

[2. Conditions de validation du projet](#_v13k27e7kunc)

[**PARTIE 1 : EXPLORATION DES DONNÉES**](#_yhuxm15chd8v)

[1. Recueil du dataset et contexte:](#_mpkuf9pmowpi)

[2. Observation des variables et premiers traitements](#_6bzydcgv31mh)

[*2.1 Variables quantitatives*](#_osuz22m7wzg2)

[*2.2 Variables qualitatives*](#_o26l1ja3l7r2)

[3. Traitement des valeurs manquantes](#_2m5yuvqktj9n)

[4. Revue statistique, visualisation 1 (avant sélection des variables)](#_4jp3hojzvt5n)

[*4.1 Représentation graphique de la variable CO2*](#_qg7npc14ifps)

[*4.2 Variables quantitatives*](#_qzrnjcynuvcz)

[*4.3 Variables qualitatives*](#_hbqdaayw4d0d)

[5. Sélection des variables utiles](#_cdigmm7ewhrm)

[6. Revue statistique, visualisation 2 (après sélection des variables)](#_xvd3pxpzqwwm)

[*6.1 Représentation graphique de la variable CO2*](#_5c0x1glxniad)

[*6.2 Variables quantitatives*](#_m02rr4qt1zbo)

[*6.3 Variables qualitatives*](#_orw5ajtcf1ir)

[**PARTIE 2 : MODÉLISATION**](#_6gd6l8i7mr4t)

[1. Régression:](#_ybb43y4e93fe)

[1.1 Objectif et méthode:](#_mkp93e615lj3)

[1.2 Préprocessing:](#_88m5nyih5cd)

[*1.3 Premier modèle*](#_qsq4xye09vd2)

[*1.4 Modèle affiné - Sélection de variables*](#_odkeukwvmjit)

[*1.5 Impact des catégories essence (ES) et diesel (GO) sur les rejets de CO2:*](#_21qw6jkp232e)

[1.6 Analyse des véhicules diesel:](#_dvgb6x7u8v70)

[1.7 Analyse des véhicules essence:](#_ptoik7h2jhm7)

[1.8 Conclusion](#_kfon7o3amnuj)

[2. Classification](#_oyqscwxdxccf)

[*2.1 Revue des variables d'intérêt*](#_m1prg1j8urcu)

[2.2 Classifieur 1 : SVM](#_gm78o6b4fm74)

[2.3 Classifieur 2 : KNN](#_jnlynqu8jm43)

[2.4 Classifieur 3 : Random Forest](#_b4r05nqmu1y9)

[*2.5 Comparaison des classifieurs : Voting Classifier et Matrices de confusion*](#_dbh4iw6ga52q)

[2.6 Interprétabilité](#_526m6b2jo3u2)

[CONCLUSION](#_5umfkh6sii91)

# **PRÉSENTATION DU PROJET**

## 1. Description

Identifier les véhicules qui émettent le plus de CO2 est important pour identifier les caractéristiques techniques qui jouent un rôle dans la pollution. Prédire à l’avance cette pollution permet de prévenir dans le cas de l’apparition de nouveaux types de véhicules (nouvelles séries de voitures par exemple).

## 2. Conditions de validation du projet

1. Chargement et exploration exhaustif des données
2. Analyse statistique et visualisations.
3. Machine learning :
   1. Prédiction du CO2
   2. Interprétabilité du modèle (SHAP)
   3. Demo Streamlit

# PARTIE 1 : EXPLORATION DES DONNÉES

## 1. Recueil du dataset et contexte:

Le jeu de données regroupe l’ensemble des caractéristiques techniques des véhicules commercialisés en France en 2013, ainsi que les consommations de carburant, les émissions de CO2 et les émissions de polluants dans l’air. Il est accessible sur le site : [Emissions de CO2 et de polluants des véhicules commercialisés en France - data.gouv.fr](https://www.data.gouv.fr/fr/datasets/emissions-de-co2-et-de-polluants-des-vehicules-commercialises-en-france/#_)

Les données recueillies comprennent **44850 enregistrements répartis sur 26 variables :**

* 13 variables numériques
* 13 variables catégorielles
* 619 doublons identifiés

En 2013, les rejets de CO2 des véhicules étaient mesurés selon le cycle d'homologation des véhicules neufs NEDC (New European Driving Cycle) décomposé en 2 cycles (cycle urbain et extra-urbain). Ce cycle définit les conditions dans lesquelles un modèle est testé (vitesse, accélération, température, utilisation limitée des rapports de boite de vitesse,...). Les véhicules étaient placés sur un banc d’essai et la mesure du CO2, des autres gaz et des polluants s'effectuait par recueil des gaz d’échappement. Ce cycle permettait également d’évaluer les consommations des véhicules.

## 2. Observation des variables et premiers traitements

La première étape de l’exploration consiste à observer les variables présentes dans le dataset dans le but de :

* les comprendre
* supprimer des informations qui semblent non pertinentes pour la suite
* identifier les valeurs manquantes

### 2.1 Variables quantitatives

Le dataset contient 13 variables quantitatives :

* puissance adm : puissance administrative du véhicule en kW
* puiss\_max : puissance maximale du véhicule en kW
* conso\_urb : consommation urbaine de carburant (en l/100km),
* conso\_extra\_urb : consommation extra urbaine de carburant (en l/100km),
* conso\_mixte : consommation mixte de carburant (en l/100km),
* CO2 : l'émission de CO2 (en g/km) qui sera la variable cible
* CO2\_type\_1 : le résultat d’essai de CO type I
* HC : les résultats d’essai HC
* NOX : les résultats d’essai NOx
* hcnox : les résultats d’essai HC+NOX
* particules : le résultat d’essai de particules
* masse\_ordma\_min : la masse en ordre de marche mini
* masse\_ordma\_max : la masse en ordre de marche maxi

**Variable cible identifiée : CO2 (g/km)**

Cette variable est continue, mais elle peut être discrétisée pour tenter de réaliser une classification par la suite. Une nouvelle variable **Cat\_CO2** est donc créée sur les base des normes suivantes :



### 2.2 Variables qualitatives

Le dataset contient 13 variables qualitatives :

* Marque : marque du véhicule
* modele\_dossier : modèle
* modèle\_UTAC : modèle UTAC (nécessité de comprendre la différence avec modèle dossier)
* design\_comm : désignation commerciale
* CNIT : Code National d'identification du type
* TVV : Type Variante Version ou type Mines
* Carburant : type de carburant
* Hybride : information permettant d'identifier les véhicules hybrides
* boite0: type de boîte de vitesse et le nombre de rapports
* champ\_V9 : champ V9 du certificat d'immatriculation qui contient la norme euro
* date\_maj : la date de la dernière mise à jour
* Carrosserie : Carrosserie
* gamme : gamme

|  |  |
| --- | --- |
| **Variable** | **Traitement** |
| Marque | Sur-représentation très forte de la marque Mercedes Benz |
| modele\_dossier | Les 2 variables semblent porter la même information. Aucun manquant dans ces variables. On ne garde que **modele\_dossier** arbitrairement pour la suite de l’étude. |
| modele\_UTAC |
| design\_comm | Ces 3 informations peuvent être retrouvées sur une carte grise de véhicule. La **désignation commerciale** est une version plus fournie que le **modèle dossier**. Il n'apportera pas à priori d'information supplémentaire dans le dataset.  Pour chaque véhicule, le constructeur doit définir un Type Variante Version (TVV) que nous retrouvons dans notre base. Le centre national de réception des véhicules associera à ce TVV un CNIT (Code National d’identification du Type). Ces deux données sont donc obligatoirement liées.  Le CNIT est généralement composé de 12 à 15 caractères et peut être partiellement décodé. il contient:   * la catégorie du véhicule * la marque du véhicule * le type d’énergie * la transmission * la tranche de puissance * le code carrosserie   Dans le cadre de notre étude, conserver ces éléments n’est pas pertinent : ils ne font que reprendre d’autres caractéristiques présentes dans d’autres variables présentes. On pourrait en revanche utiliser les données présentes dans ces codes pour remplacer d’éventuelles données manquantes.  Ces 3 variables sont donc supprimées de notre étude. |
| CNIT |
| TVV |
| Carburant | Recodage nécessaire de la variable:  ES = Essence  GO = Gasoil  EH = Essence Hybride Non Rechargeable  EE = Essence Hybride Rechargeable  EL = Electrique  GH = Gasoil Hybride Non Rechargeable  ES/GP = GP/ES = Essence GPL  ES/GN = GN/ES = Essence Gaz Naturel  FE = Superethanol-E85  GN = Gaz Naturel (Vehicule)  GL = Gasoil Hybride Rechargeable  On constate une sur-représentation de la modalité Gasoil (GO) dans le dataset. |
| Hybride | Variable à 2 modalités : oui/non |
| boite0 | Type de boîte de vitesse et nombre de rapports. Cette variable est séparée pour obtenir ces 2 informations dans des variables distinctes:   * une variable **boîte** (Manuelle ou Automatique) * une variable **rapport** (de 1 à 6) |
| champ\_V9 | Champ de la carte grise qui procure différentes informations comme la date de commercialisation, ou la norme euro. De cette variable nous ne récupérons que la norme euro pour la suite de notre étude |
| date\_màj | date de mise à jour de la base |
| Carrosserie | MINIBUS 32576  BERLINE 4651  BREAK 2219  TS TERRAINS/CHEMINS 1145  COUPE 1102  COMBISPACE 880  MONOSPACE COMPACT 608  CABRIOLET 607  MINISPACE 145  MONOSPACE 114  On constate une forte sur-représentation de la catégorie ‘minibus’ |
| gamme | MOY-INFERIEURE 20138  MOY-SUPERIEURE 15003  LUXE 5151  SUPERIEURE 1954  INFERIEURE 1606  ECONOMIQUE 195 |

Une fois toutes les variables observées, il faut traiter les valeurs manquantes.

## 3. Traitement des valeurs manquantes

Les valeurs manquantes concernent les variables suivantes:

conso\_urb 10

conso\_extra\_urb 10

conso\_mixte 10

CO2 10

CO2\_type\_1 106

HC 34 132

NOX 106

hcnox 9 974

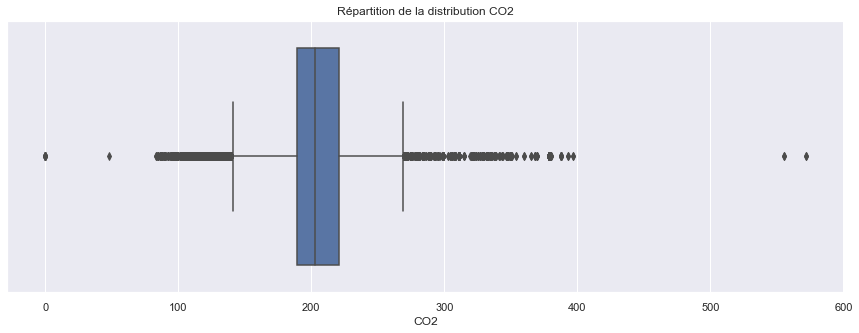
particules 2 561

|  |  |  |
| --- | --- | --- |
| **Variable** | **Nombre de valeurs manquantes** | **Traitement** |
| conso\_urb | 10 | Correspondent aux véhicules électriques, qui n’émettent pas de CO2. Les valeurs manquantes sont donc remplacées par 0. |
| conso\_extra\_urb |
| conso\_mixte |
| CO2 |
| CO2\_type\_1 | 106 | Remplacement par les moyennes de la variable sur les données renseignées sauf pour la variable HC qui a été supprimée car elle avait trop de valeurs manquantes. |
| HC | 34 132 |
| NOX | 106 |
| hcnox | 9 974 |
| particules | 2 561 |

## 4. Revue statistique, visualisation 1 (avant sélection des variables)

### 4.1 Représentation graphique de la variable CO2

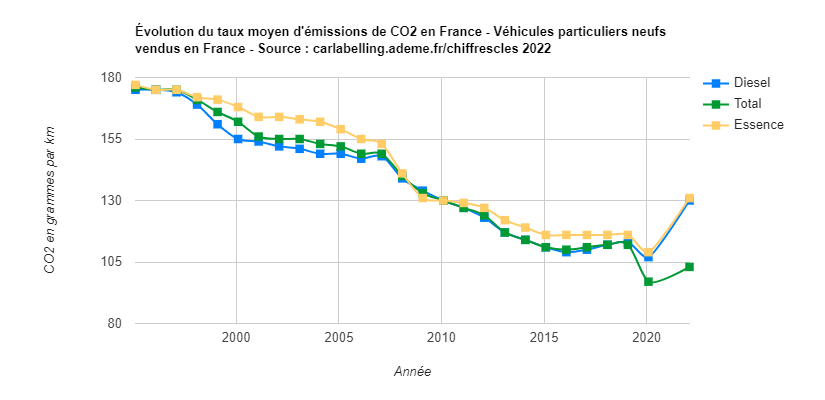
Une boîte à moustache est utilisée pour représenter la répartition de la variable CO2:



On peut déjà observer des outliers :

* autour de 0 : il s’agit des véhicules électriques
* 2 véhicules à très fort taux de CO2 (à plus de 550 g/km)

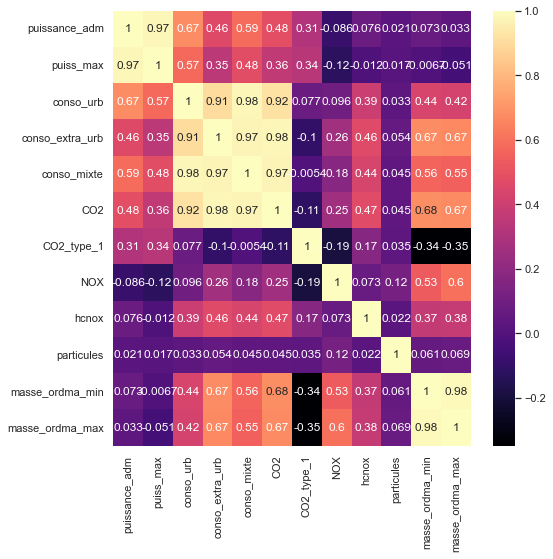
L’émission médiane des véhicules présents dans notre dataset est autour de 200 g/km. Si on se réfère aux données françaises, on est au-dessus de ce qui est constaté :



Bien que la base de données exploitée date de 2013, notre point médian est au-dessus de ce qui est normalement constaté.

### 4.2 Variables quantitatives

Une fois les variables observées, la cible identifiée et les données nettoyées, nous pouvons nous intéresser aux corrélations entre les variables numériques de notre dataset. Grâce à une heatmap, nous obtenons les résultats ci-dessous.



Cette heatmap montre:

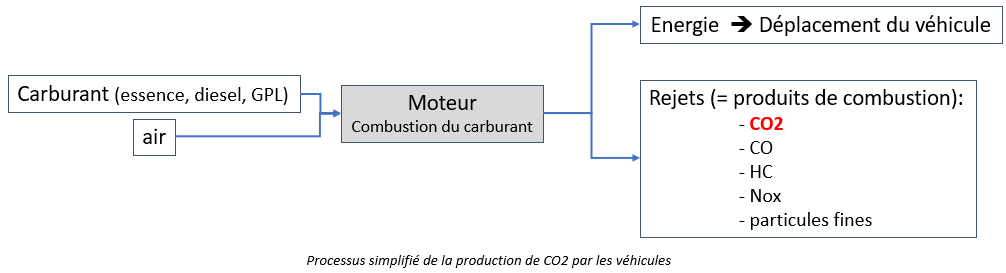
* une colinéarité entre les deux variables liées à la masse du véhicule (min et max),
* une forte corrélation positive entre les différents types de consommation entre eux et avec les émissions de CO2.

→ La variable masse\_ordma\_max a été supprimée et la variable masse\_ordma\_min conservée.

→ Pour les autres variables, l’approche métier nous montre que:

* les rejets de CO2 sont directement liés à la consommation de carburant. L’origine du CO2 provient du carburant, pour l’atome de carbone et de l’oxygène de l’air pour les 2 atomes d’oxygène,
* comme le CO2, les autres gaz (CO, HC, NOX) et les particules sont des produits de réaction chimique issus de la combustion du carburant dans le moteur.

Ceci est illustré par le schéma suivant:



On en déduit:

1. le calcul de rejet de CO2 par type de carburant (en prenant en compte chimiquement les masses molaires d'octane, les masses d'eau, les masses de CO2 et la masse volumique du carburant ):

* un litre d'essence rejette 2.3 kg de CO2,
* un litre de diesel rejette 2.6 kg de CO2,
* un litre de GPL rejette 1,7 kg de CO2.

Les rejets de CO2 correspondent à une réaction chimique que l’on peut simplifier par l’équation suivante:

Consommation (l/km) \* constante = CO2 (g/km)

Les rejets de CO2 sont directement liés à la consommation de carburant.

Aucune variable de consommation ne sera donc conservée (les prendre en compte reviendrait à mettre la variable CO2 (= objectif) dans les variables explicatives).

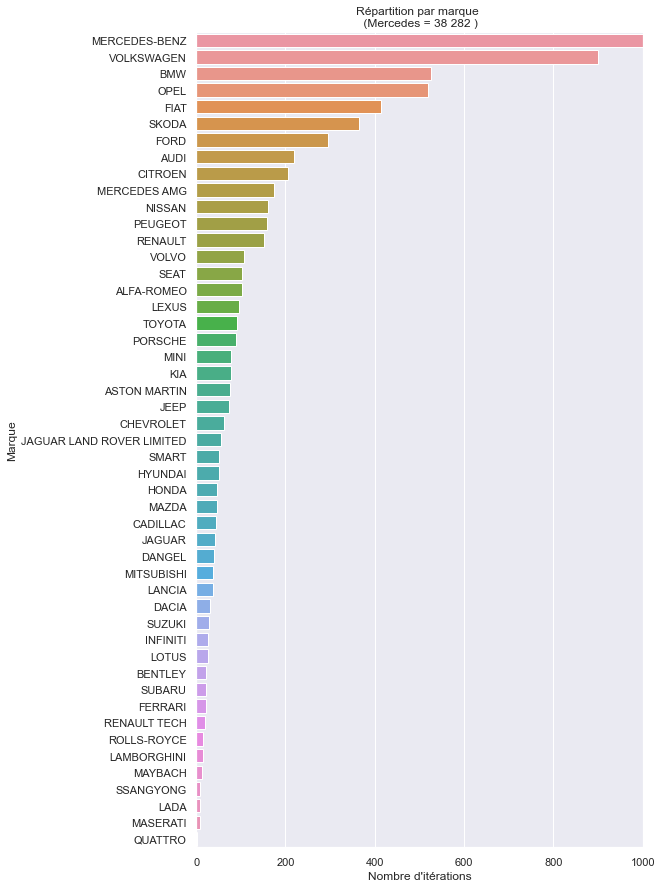
1. l’absence de relation de causalité entre le CO2 et tous les autres rejets. Tous ces rejets sont des produits de combustion du carburant et n’ont donc pas de lien de causalité avec le niveau de rejet de CO2. Le CO2 est un gaz à effet de serre influençant le climat. Les autres rejets sont des polluants impactant la santé.

De plus, cette étude consiste à déterminer les caractéristiques techniques influençant les rejets de CO2. Ces autres rejets n’en sont pas.

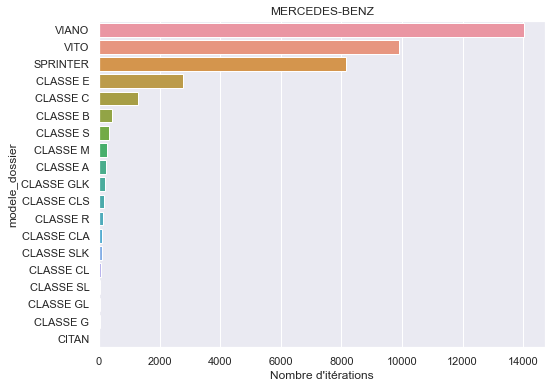
Aucun de ces paramètres ne sera donc conservé dans la poursuite de cette étude.

### 4.3 Variables qualitatives

En observant la marque des véhicules, on retrouve la prépondérance de la marque Mercedes-Benz, mais également un grand nombre de véhicules Volkswagen :



En zoomant sur les modèles Mercedes-Benz présents, on obtient la répartition suivante:



Il s’agit de modèles type fourgons qui peuvent expliquer le nombre de “minibus” constatés lors de la première observation de la base.

Afin d’établir s’il existe un lien statistique significatif entre les émissions de CO2 et les variables qualitatives, des ANOVA sont réalisées. Les résultats sont les suivants:

|  |  |
| --- | --- |
| **Variable** | **Résultat de l’ANOVA** |
| Marque |  |
| modele\_dossier |  |
|
| Carburant | Les carburants Essence et Gasoil représentent 99,3 % du portefeuille : on ne conserve que ces 2 modalités. |
| Hybride | Les véhicules non-hybrides représentent 98% de tous les véhicules du dataset. Les véhicules hybrides ne sont pas suffisamment représenté |
| boite0 |  |
| Carrosserie |  |
| gamme |  |

Les ANOVA réalisées montrent tous un lien significatif entre les variables sélectionnées et la variable à expliquer CO2. Cette étape d’observation permet cependant de repérer des problèmes de déséquilibre sur certaines variables.

## 5. Sélection des variables utiles

Les étapes précédentes ont permis de :

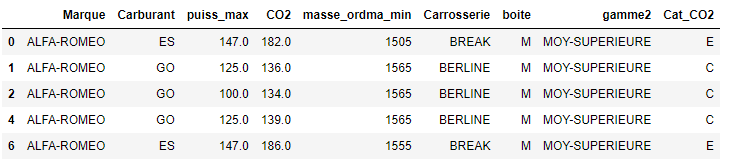
* supprimer les valeurs manquantes
* effectuer un premier tri des variables à sélectionner
* identifier les variables représentant un lien significatif avec les émissions de CO2
* repérer les problématiques de déséquilibres

Suite à ces différentes étapes, les variables sélectionnées pour effectuer une modélisation sont les suivantes:

|  |  |
| --- | --- |
| Variables quantitatives | Variables qualitatives |
|  |  |
| puiss\_max | Marque |
| masse\_ordma\_min | Carburant |
|  | Carrosserie |
|  | boite |
|  | gamme2 |

Les variables qui seront expliquées sont la variable **CO2 (pour la régression)** et **Cat\_CO2 (pour la classification)**

Le dataset ainsi retravaillé ressemble à :

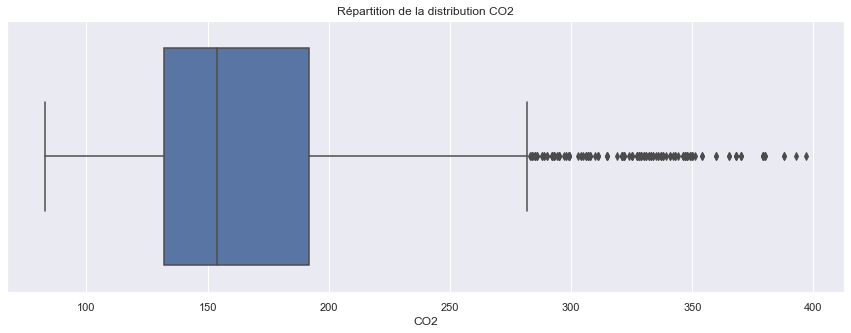


Ce tri dans les variables a un effet de générer à nouveau des doublons. La suppression de ces doublons réduit considérablement la taille de la base. Au départ, elle contenait 43 728 lignes, après sélection des variables et dédoublonnage, elle en contient 5020. Une deuxième passe de revue statistique est nécessaire afin d’observer les effets de ces actions.

## 6. Revue statistique, visualisation 2 (après sélection des variables)

### 6.1 Représentation graphique de la variable CO2

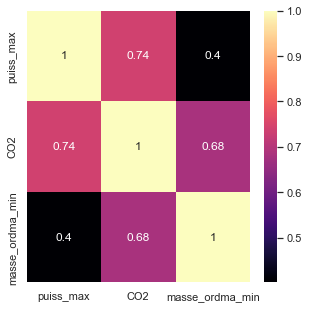
En premier lieu, on reprend la variable CO2:



On observe une baisse conséquente de la médiane des émissions de CO2. La baisse est évaluée à 50 g/ km.

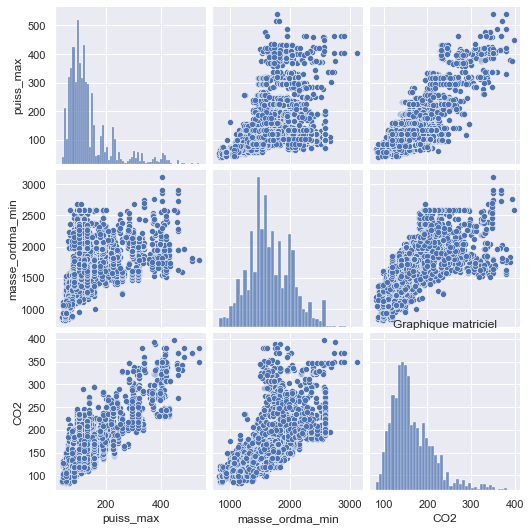
### 6.2 Variables quantitatives

***Heatmap***



Une fois la sélection des variables effectuée, la heatmap obtenue confirme bien l’impact de la puissance et de la masse du véhicule sur les rejets de CO2. La suite de l’étude va nous permettre de détailler ces corrélations.

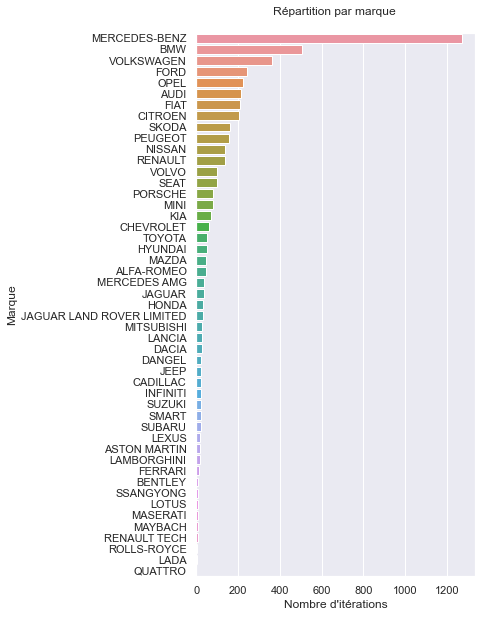
***Graphique matriciel***



Le pairplot nous permet de constater qu’il n’existe pas de fortes corrélations entre nos variables quantitatives, donc pas de colinéarité des variables explicatives sélectionnées.

### 6.3 Variables qualitatives

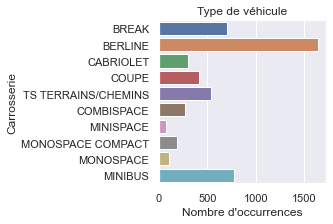
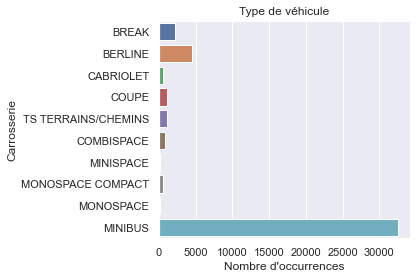
Les variables qualitatives ainsi sélectionnées offrent un dataset différent du premier étudié :



La marque Mercedes Benz est toujours sur-représentée, mais dans des proportions plus raisonnables qu’au départ.

On observe également que la modalité “Minibus” est revenue dans des proportions plus cohérentes :

***Avant Après***



Des ANOVA sont refaites pour tester si le lien est toujours significatif entre les émissions de CO2 et chacune des variables: toutes les ANOVA réalisées donnent bien ce lien significatif après notre épuration de la base de données.

# PARTIE 2 : MODÉLISATION

## 1. Régression:

### 1.1 Objectif et méthode:

Le dataset est composé d’une variable cible numérique et de variables explicatives numériques et catégorielles. Il est donc possible d’effectuer une régression multiple comme première approche pour déterminer quelles sont les variables responsables des variations des rejets de CO2.

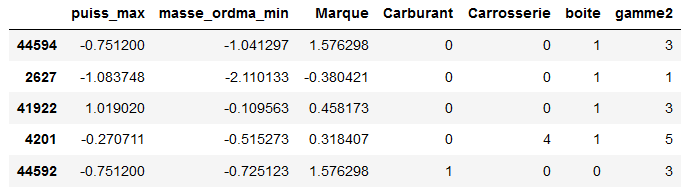
La méthode choisie consiste à effectuer une régression multiple sur l’entièreté du dataset, puis une deuxième régression est effectuée après sélection des variables les plus influentes. Ce principe est répété en réalisant des régressions sur des catégories plus précises du dataset afin d'améliorer la performance des modèles.

### 1.2 Préprocessing:

Afin d’effectuer la modélisation, une étape de preprocessing:

* numérise les variables catégorielles grâce à la fonction LabelEncoder(),
* standardise, grâce à la fonction StandardScaler():
  + les variables numériques,
  + les variables catégorielles numérisées, si le nombre de catégories est supérieur à 10.

Les variables sont ainsi regroupées dans le dataset suivant :



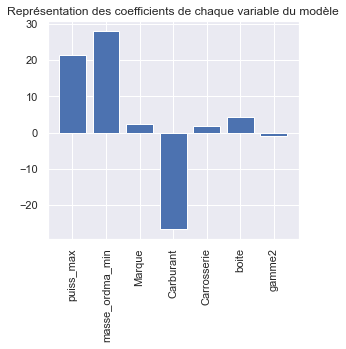
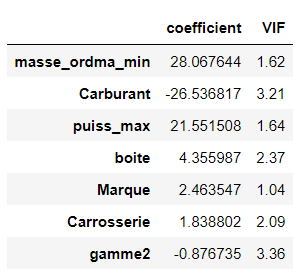
Ce dernier dataset est divisé en un jeu d'entraînement (80%) et un jeu de test (20%) avec un random\_state à 123 pour garantir la reproductibilité des résultats (à noter que, lors de la réalisation de l’étude, ce dernier argument n’était pas renseigné, permettant ainsi de tester la robustesse des modèles).

### 1.3 Premier modèle

Après instanciation d’un modèle de régression linéaire (LinearRegression()), ce dernier est ajusté sur le jeu d’entraînement et évalué par validation croisée. Enfin, les metrics R², MAE et RMSE seront calculés sur les 2 jeux et synthétisés dans le tableau suivant:

|  |  |  |  |
| --- | --- | --- | --- |
| **Metrics** | *Entraînement* | *Test* | *Validation croisée* |
| **R²** | 0.78 | 0.78 | 0.78 |
| **MAE** | 17.88 | 18.39 | - |
| **RMSE** | 23.26 | 23.59 | - |

Les scores obtenus sur l'échantillon d'apprentissage, par validation croisée et sur l'échantillon de test sont corrects et ne montrent pas de sur-apprentissage. On note une dispersion importante. Le modèle parvient à apprendre sur les données d'entraînement et à généraliser sur les données de test avec une erreur de 11% en moyenne.

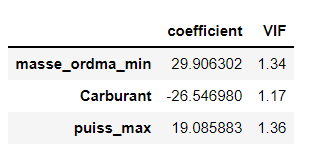
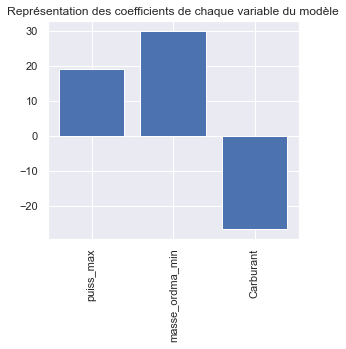


Pas de colinéarité entre variables. Il est possible d'affiner le modèle.

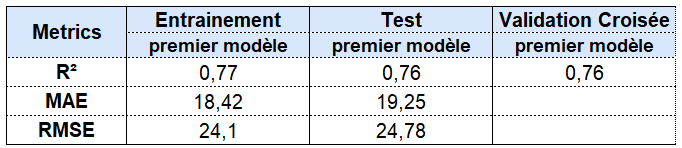
### 1.4 Modèle affiné - Sélection de variables

Après sélection des variables à l’aide du sélecteur SelectFromModel(lr), 3 variables sont sélectionnées sur les 7 initiales (pas de colinéarité détectée):

* la puissance maximale,
* la masse minimum
* le carburant.



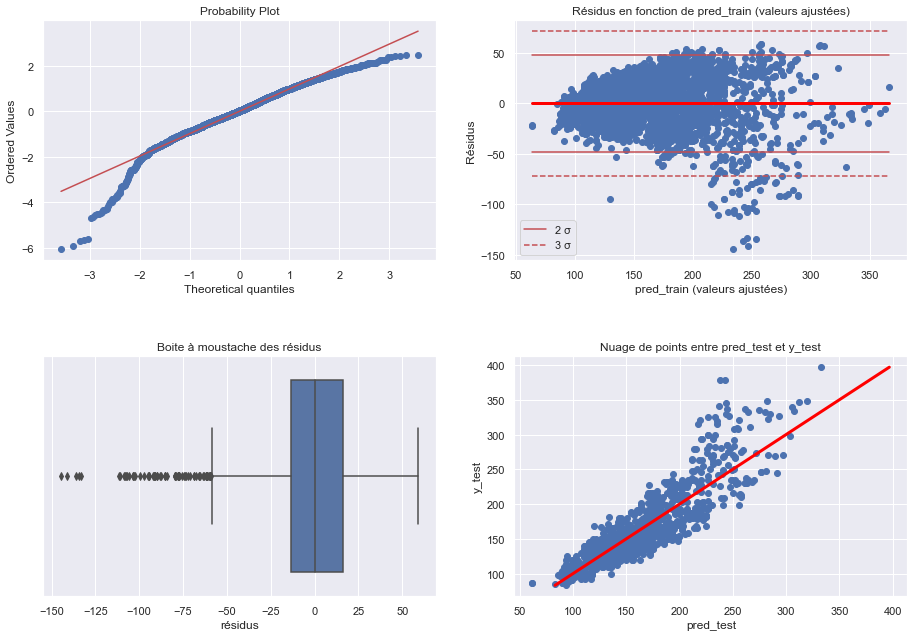
Les metrics sont calculés et synthétisés dans le tableau suivant:



Les résultats sont presque identiques au premier modèle. On ne remarque que très peu de perte d'informations (les R² sont bons et proches du premier modèle) alors que ce dernier modèle possède 2 fois moins de variables explicatives. On note une légère augmentation de la dispersion.

Le modèle parvient à apprendre sur les données d'entraînement et à généraliser sur les données de test avec **une erreur de 12% en moyenne.**

Une analyse des résidus fournit les graphes suivants :



**Graphes Probability plot: Résidus en fonction des pred\_train :** Graphiquement, les résidus ne suivent pas une loi normale (confirmé par la p\_value du test de normalité jarque-bera à 0). On remarque beaucoup de points extrêmes. La présence de groupes pourrait en partie expliquer ce phénomène.

**Résidus en fonction des pred\_train :** On note un profil particulier d’entonnoir avec beaucoup de résidus à moins de - 3 écarts-types.

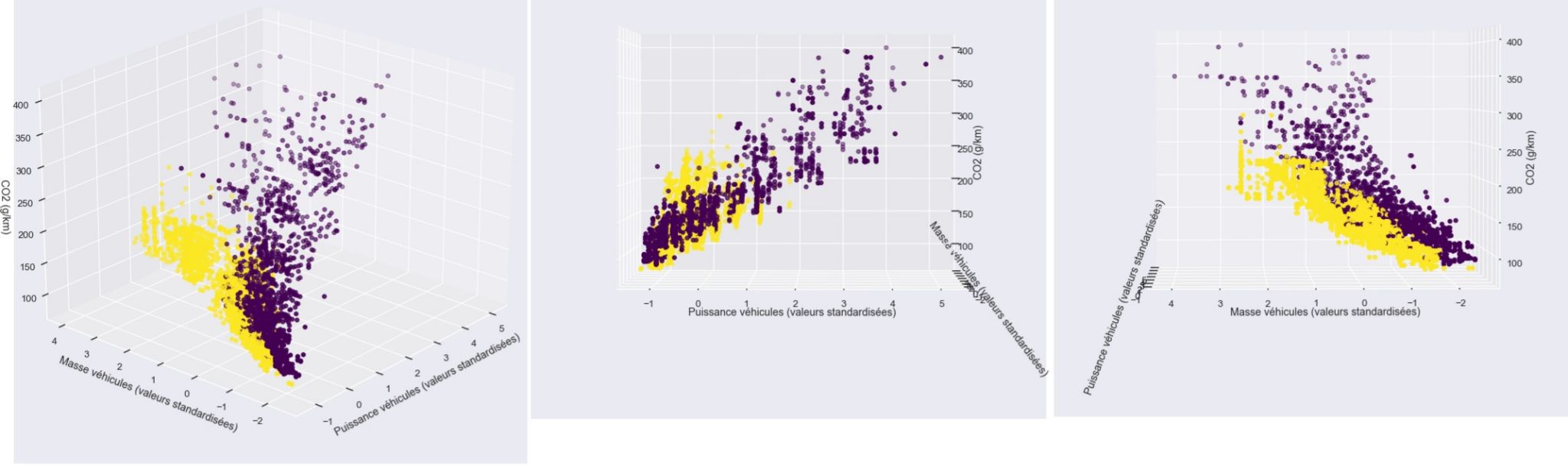
**Graphe Nuage de points entre pred\_test et y\_test :** On observe une relation avec un léger aspect d' “entonnoir asymétrique" :

* plus y\_test augmente, plus les points s'éloignent de pred\_test (la dispersion augmente avec y\_pred),
* à partir d'une prévision d'environ 210 g/km de CO2, la dispersion est beaucoup plus importante surtout d'un côté : les rejets réels des véhicules sont beaucoup plus importants que prévu.

On constate qu’on a trop de points aberrants et éloignés.

***Remarque :*** dans le dataset originel, 2 points aberrants apparaissaient. Ces véhicules rejetaient nettement plus de CO2 que les autres (plus de 500g de CO2 / km). Ces véhicules ont été supprimés lors de la préparation du dataset ([*cf. Suppression des outliers*](https://docs.google.com/document/d/1HCexRAW7Tgpmp7v7nLWljxJokRpBKeE8fzF79pEx2Q8/edit)).

### 1.5 Impact des catégories essence (ES) et diesel (GO) sur les rejets de CO2:



Représentation des rejets de CO2 des véhicules en fonction de leurs masses, leurs poids et leurs carburants

(Les points violets représentent les véhicules Essence et les points jaunes les véhicules Diesel)

La variable carburant est statistiquement significative, c'est-à-dire qu'on observe une différence significative entre les véhicules essence et diesel. Ceci semble cohérent. L'essence et le diesel ont des propriétés différentes et les moteurs associés ont des technologies différentes.

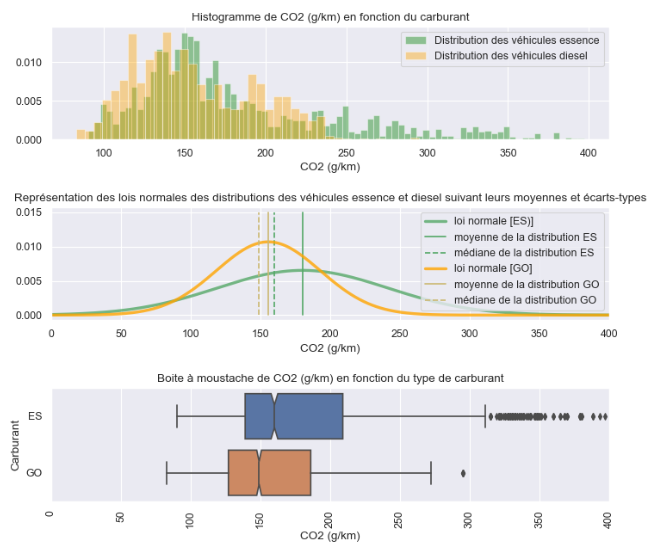
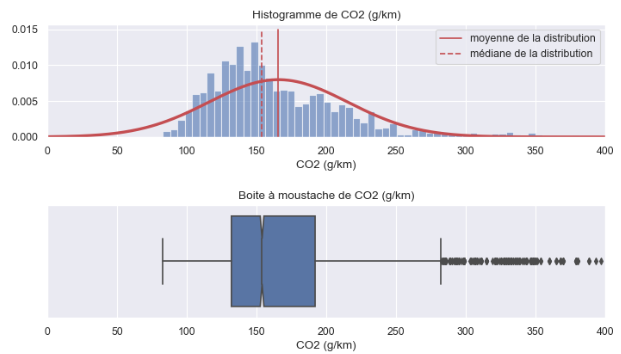
Le graphe précédent nous montre bien l'impact de la variable carburant sur les rejets de CO2. Il semble que les rejets de CO2 sont plus dépendants de la masse du véhicule pour les véhicules diesel (points jaunes) et de la puissance maximale pour les véhicules essence (points violets).

Il pourrait donc être judicieux de créer 2 modèles différents (un essence et un diesel) pour le confirmer.

En regardant les graphes, on constate que les distributions essence et diesel sont décalées. La distribution des véhicules essence est plus étalée et ils concentrent la majorité des points extrêmes.

On note un aspect de distribution plus asymétrique pour les véhicules essence et un aspect bimodal pour les véhicules diesel.

De manière générale, les véhicules diesel rejettent moins de CO2 que les véhicules essence.

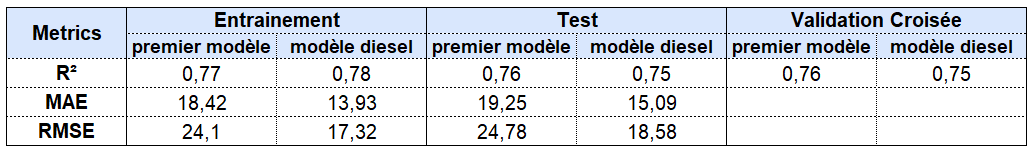


Ces 2 distributions (essence et diesel) ont été analysées selon la même méthode. Seuls les résultats des modèles affinés sont présentés.

### 1.6 Analyse des véhicules diesel:

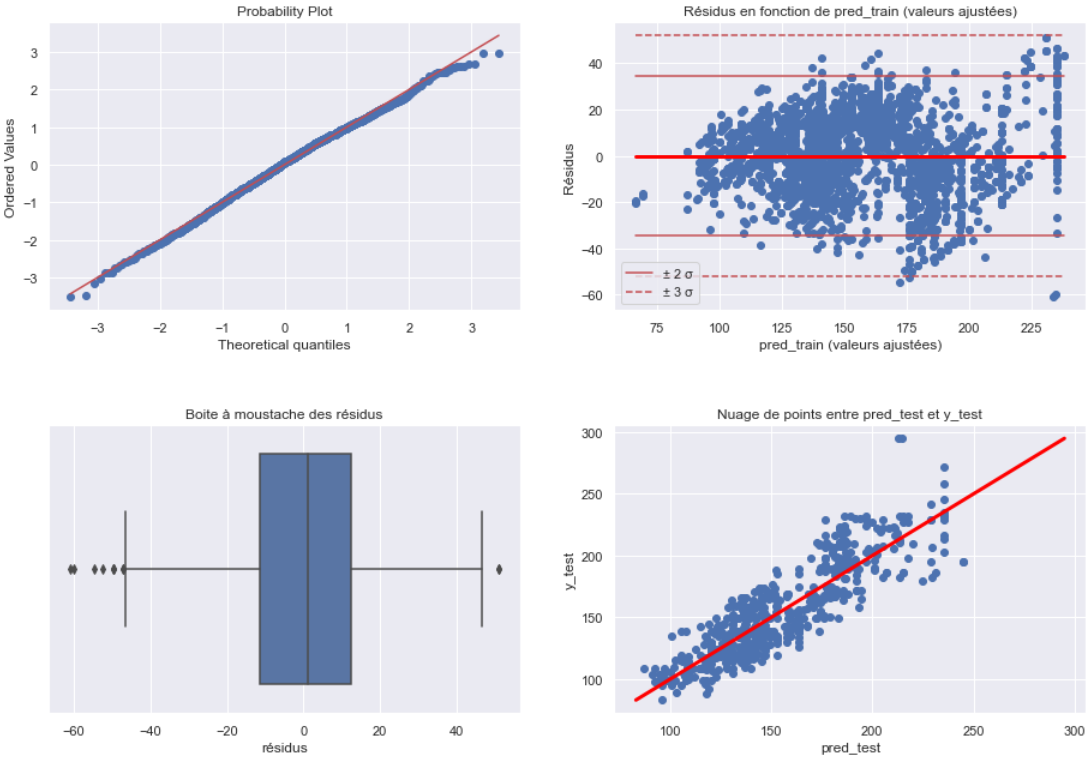
La sélection de variables influentes montre que les rejets de CO2 des véhicules diesel ne dépendent principalement que d’une seule variable, la masse minimum.

Les métriques de ce modèle affiné sont comparés à ceux du premier modèle dans le tableau suivant :



Les R² du modèle diesel affiné sont bons et identiques au modèle général. Pas de perte d’information.

Les MAE et RMSE du modèle diesel affiné sont plus faibles. La dispersion est réduite. Cette dernière modélisation est donc meilleure que la modélisation précédente.



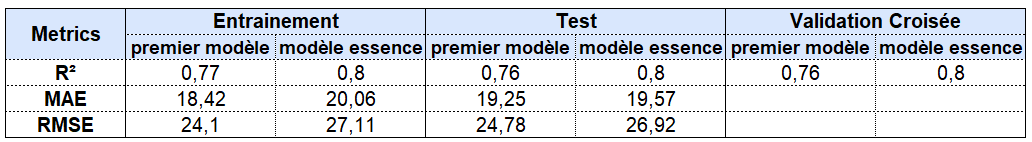
La représentation des résidus du modèle diesel affiné montre que les résidus semblent être normalement distribués. La distribution est moins dispersée que précédemment (confirmant que la dispersion de la distribution générale est causée majoritairement par les véhicules essence) et le nombre de points extrêmes a nettement diminué.

### 1.7 Analyse des véhicules essence:

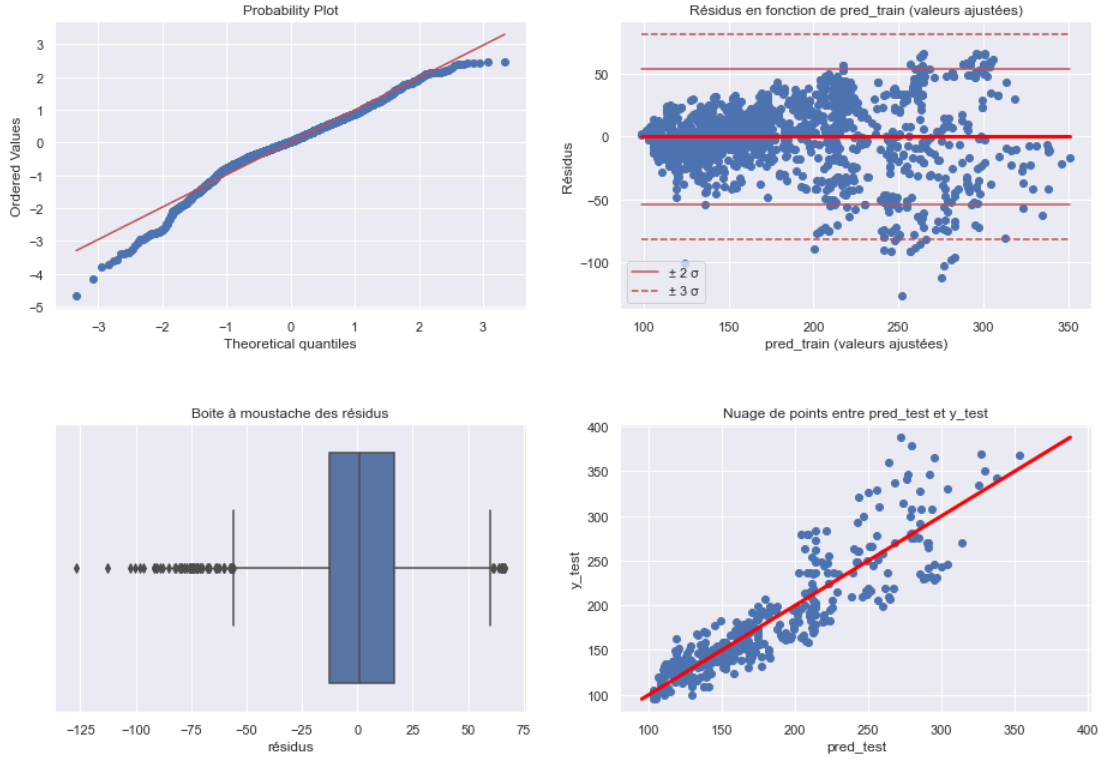
La sélection de variables influentes montre que les rejets de CO2 des véhicules essence ne dépendent principalement que de 2 variables:

* la puissance maximale,
* la masse minimale.

Les métriques de ce modèle affiné sont comparés à ceux du premier modèle dans le tableau suivant :



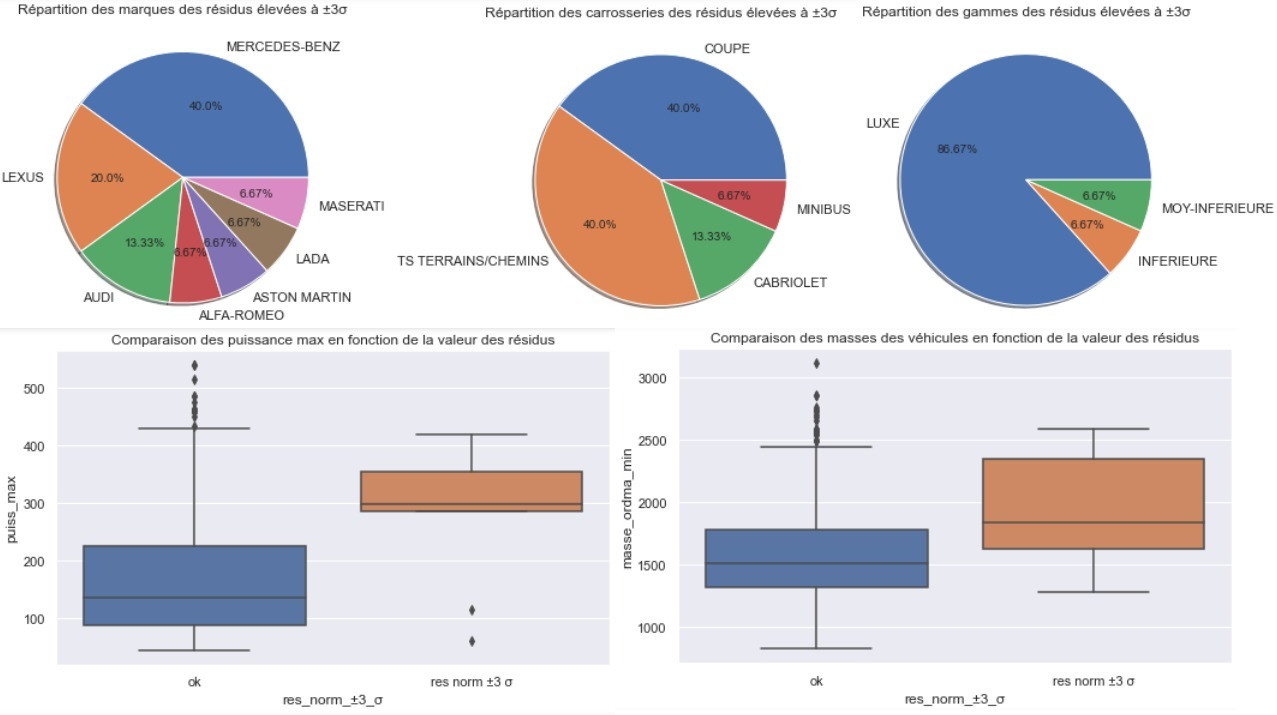
Les R² sont meilleurs aussi bien sur le jeu d'entraînement que sur le jeu de test. Cependant, la qualité du modèle s’est dégradée (MAE et RMSE plus grandes).



Le modèle ne suit pas une loi normale (p\_value = 0). Le graphique probabilty\_plot montre bien une dérive sur les 2 extrémités. On note une répartition des résidus en forme d'entonnoir. Plus les rejets de CO2 augmentent, plus la dispersion augmente, plus le nombre de véhicules diminue. Les résidus à plus ou moins 3 écarts-types semblent nombreux et nécessitent une attention particulière. En effet, 0.9% de points sont à plus ou moins 3 écarts-types (0.3% attendus si la distribution suit une loi normale). On remarque également que ces points influencent grandement le modèle (distance de Cook élevée):



Après avoir enrichi le dataset initial des valeurs des résidus, la sélection des véhicules essence à plus ou moins 3 écarts-types montrent:



Ces véhicules ont des caractéristiques bien particulières pouvant expliquer leur décalage par rapport au modèle. Il s’agit de véhicules:

* de luxe (+ 80%), majoritairement de type coupés ou type tout terrain (80%), de marques facilement identifiables pour ces types de véhicules (Mercedes-Benz, Lexus, Maserati, Aston Martin)
* significativement plus puissants et plus lourds que la majorité des autres véhicules.

### 1.8 Conclusion

Les analyses de régression nous montrent que les caractéristiques techniques responsables des rejets de CO2 des véhicules sont

* la masse pour les véhicules diesel,
* la masse et la puissance maximale pour les véhicules essence.

De manière générale, les véhicules diesel rejettent moins de CO2/km que les véhicules essence.

Plus la puissance et/ou la masse du véhicule augmentent, plus ce véhicule rejettera de CO2.

## 2. Classification

L’objectif ici est de construire des modèles permettant de classifier les véhicules selon les normes européennes en fonction de leurs émissions en CO2.

Afin d'effectuer ces classifications, nous avons sélectionné les 3 types de classifieurs suivants qui sont adaptés à des problématiques de classifications “multi-classes” :

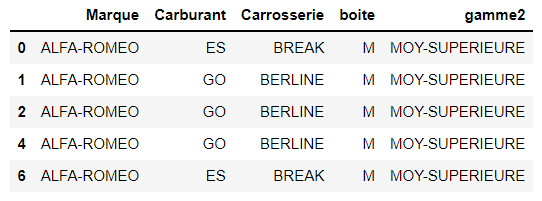
* SVM : les séparateurs à vaste marge
* KNN : les K plus proches voisins (ou K-nearest neighbors)
* Random Forest : les forêts d’arbres décisionnels

Pour chaque classifieur, un modèle sans hyperparamètre est instancié, puis un second modèle est instancié avec recherche des meilleurs hyperparamètres (GridSearchCV()). Le choix du meilleur classifieur est fait en comparant les accuracy et les f1\_score de chaque classifieur entre eux.

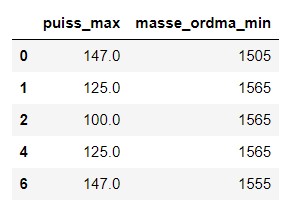
### 2.1 Revue des variables d'intérêt

La classification tient compte des variables sélectionnées après l’analyse des données.

* Les variables catégorielles



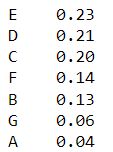
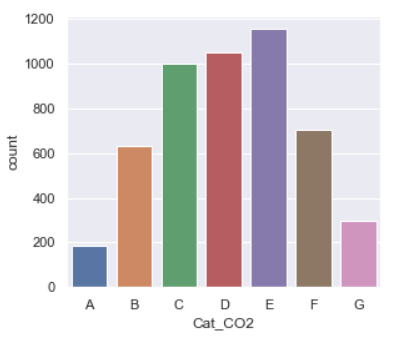
* Les variables numériques



Afin de pouvoir utiliser ces classifieurs, nous avons :

* transformé les variables catégorielles en indicatrices
* standardisé les variables numériques

Notre variable cible est la variable **Cat\_CO2.** On vérifie la répartition des classes pour évaluer l’équilibre.



La répartition est légèrement déséquilibrée avec une faible présence de voitures des classes A et G (respectivement 4% et 6% du dataset).

### 

### 2.2 Classifieur 1 : SVM

On instancie un classifieur *clf\_svc* à partir de la fonction SVC en définissant uniquement l’hyperparamètre *gamma* :

clf\_svc = SVC(gamma = 'scale')

|  |  |
| --- | --- |
| **Classifieur initial** | **Classifieur optimisé** avec les hyperparamètres suivants :   * la constante C : 200 * le kernel : ‘rbf’ * le coefficient gamma : 0.1 |
| accuracy score : 71 % | accuracy score : 74 % |

Au regard du f1-score, le classifieur SVM est globalement satisfaisant. On note toutefois qu’il a beaucoup de mal à classer les voitures de la classe A (peu polluantes).

### 

### 2.3 Classifieur 2 : KNN

On instancie un classifieur *clf\_knn* sans aucun hyperparamètre:

clf\_knn = KNeighborsClassifier()

|  |  |
| --- | --- |
| **Classifieur initial** | **Classifieur optimisé** avec les hyperparamètres suivants :   * n\_neighbors : 1 * metric : ‘minkowski’ * leaf\_size : 1 * p : 1 |
| accuracy score : 61 % | accuracy score : 70 % |

En regardant le f1-score, le classifieur KNN est intéressant pour les classes de voitures les plus polluantes E, F et G. Par contre, il est peu intéressant pour les autres classes où le f1-score est en dessous de 70%.

### 

### 2.4 Classifieur 3 : Random Forest

On instancie un classifieur *clf\_rf* sans aucun hyperparamètre:

clf\_rf = RandomForestClassifier()

|  |  |
| --- | --- |
| **Classifieur initial** | **Classifieur optimisé** avec les hyperparamètres suivants :   * n\_estimators qui est le nombre d’arbres de décisions : 400 * criterion qui mesure la qualité d’une division de l’arbre : ‘gini’ * max\_features qui définit le nombre de caractéristiques à prendre en compte afin de réaliser la meilleure répartition : 'sqrt' |
| accuracy score : 74 % | accuracy score : 74 % |

Les f1-score obtenus avec les classifieurs RF sont globalement plus équilibrés, même si on note une performance plus importante dans la classification des véhicules plus polluants.

### 2.5 Comparaison des classifieurs : Voting Classifier et Matrices de confusion

Après exécution du Voting classifier, on compare les scores des différents classifieurs :

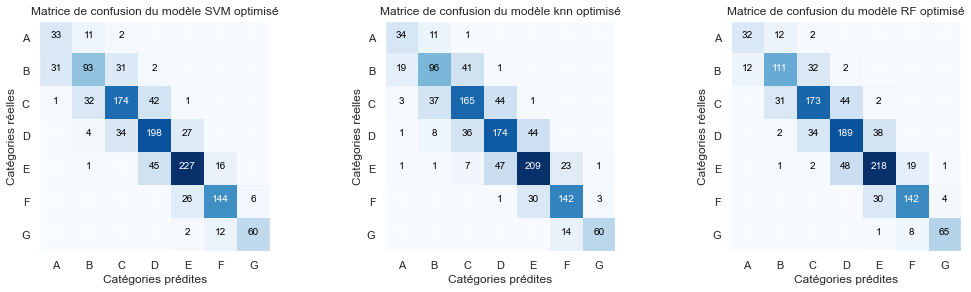
Accuracy SVM = 0.73

Accuracy KNN = 0.701

Accuracy RF = 0.749

Accuracy Voting Classifier = 0.739

La comparaison des 3 modèles optimisés nous donne les matrices de confusion suivantes:

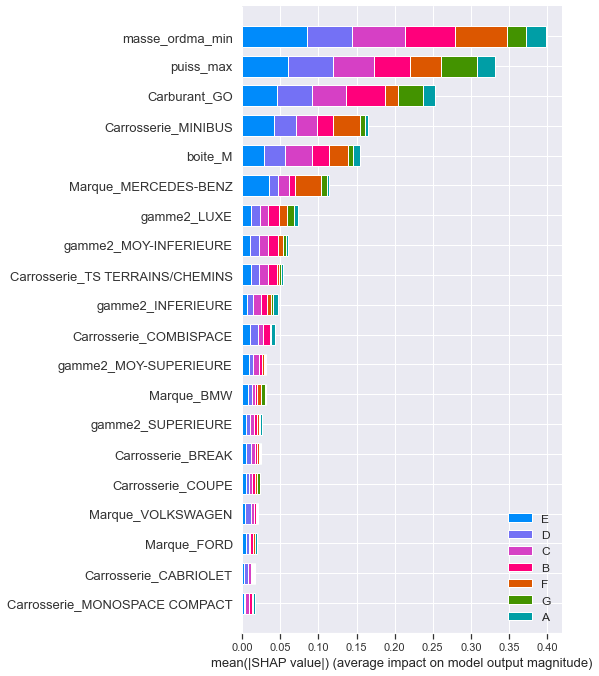


Le classifieur RF semble le plus adapté car en plus d’avoir un score de précision plus élevé, il est aussi plus homogène sur l’ensemble des classes.

# 

### 2.6 Interprétabilité

Les 3 variables les plus importantes sont la masse, la puissance et le type de carburant. Les variables semblent impacter toutes les classes mais avec des intensités différentes.



* On observe notamment que les variables ont des impacts très similaires pour les classes B, C, D et E. Il est donc plus difficile de discriminer ces classes. Ceci explique notamment le résultat obtenu dans la matrice de confusion où ces 4 classes concentrent les erreurs de classification.
* Les impacts sont plus distincts pour les classes F et G, qui sont celles où on a de meilleurs scores

On peut également afficher les summary plots par catégorie :

|  |  |
| --- | --- |
| **Catégorie A** (regroupant les véhicules les moins polluants) : les véhicules appartenant à cette classe ont une faible puissance, une faible masse, une boîte manuelle et sont de type diesel | **Catégorie B** : les 4 variables les plus influentes sont les mêmes que pour la catégorie A. Cependant, la variable carburant est plus influente que la variable puissance pour cette catégorie B. |

|  |  |
| --- | --- |
| **Catégorie C** | **Catégorie D** |

|  |  |
| --- | --- |
| **Catégorie E** | **Catégorie F** |

|  |  |
| --- | --- |
| **Catégorie G** | **Général :** les 2 variables les plus influentes sur l'ensemble des classes sont la masse et la puissance. On note que plus on avance dans les classes, plus la masse et la puissance des véhicules augmentent. On constate que pour la catégorie la plus polluante (G), la variable carburant joue un rôle majeur alors qu'elle semble jouer un rôle de moins en moins important en passant de la catégorie A à F. |

Dépendance plot. Les liens entre les variables influentes du modèle sont illustrés pour chaque classe:

**Puissance vs Boîte de vitesse**

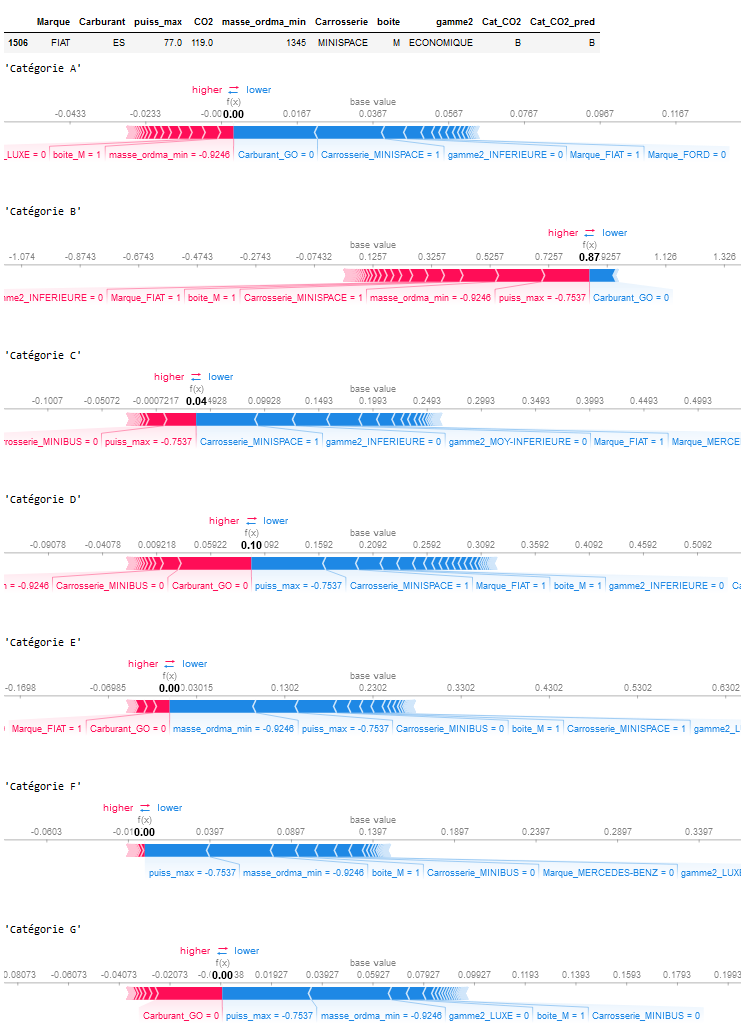
|  |  |
| --- | --- |
| **Catégorie A** | **Catégorie B** |
| **Catégorie C** | **Catégorie D** |
| **Catégorie E** | **Catégorie F** |
| **Catégorie G** | L’analyse des graphes nous montre que les voitures ayant une boîte de vitesse manuelle et une puissance faible sont moins polluantes (Catégorie A).  Lorsqu’on évolue dans les catégories, on a des voitures avec des puissances de plus en plus élevées et des boîtes automatiques (exemple des Catégories F et G). |

**Puissance vs Masse mini du véhicule**

|  |  |
| --- | --- |
| **Catégorie A** | **Catégorie B** |
| **Catégorie C** | **Catégorie D** |
| **Catégorie E** | **Catégorie F** |
| **Catégorie G** | Les véhicules de faible puissance et de faible masse sont les moins polluants (Catégorie A).  En parcourant les catégories, sont progressivement plus lourds et plus puissants (à partir de la Catégorie C). La catégorie G n’est constituée que de véhicules avec puissance et masse très élevées. |

Force plot

Affichage de tous les force\_plot (un force\_plot par catégorie de pollution) pour un véhicule sélectionné. L’exemple suivant montre un véhicule bien classé (catégorie réelle B / catégorie prédite B) :



Il y a autant de force\_plot que de classes cibles. Le modèle prédit la classe avec le score le plus élevé (ici catégorie B). On note que les variables puissance et masse ont fortement contribué à classer ce véhicule en catégorie B. A l’inverse, plus on descend dans les catégories (D vers G), plus ces variables contribuent à ne pas classer ce véhicules dans ces catégories. On remarque que c’est surtout la variable carburant qui a contribué à ne pas classer ce véhicules en catégorie A.

L’exemple suivant montre un véhicule mal classé (catégorie réelle C / catégorie prédite E) :



On remarque que les variables responsables de la catégorisation du véhicule en catégorie E sont les variables carrosserie\_TS TERRAINS/CHEMINS et gamme2\_MOY-SUPERIEURE. Or, ces variables ne sont pas les plus influentes pour bien classifier les véhicules, contrairement aux variables puissance et masse par exemple. La catégorisation de ce véhicule en catégorie E s’explique par l’influence trop forte de variables peu significatives.

# CONCLUSION

L’objectif du projet était de prédire, en fonction d’un certain nombre de caractéristiques, les émissions de CO2 de véhicules. Pour ce faire, la régression et la classification ont été utilisées et donnent des résultats satisfaisants. Les variables identifiées comme étant les plus pertinentes pour prédire les taux de CO2 sont, par ordre décroissant d’importance :

* la masse du véhicule
* la puissance du véhicule
* le type de carburant (essence ou diesel)

Le projet a montré l’importance du préprocessing des données. Ce travail a permis d’améliorer la base utilisée pour la régression/classification en rééquilibrant le dataset de départ qui comportait beaucoup de doublons (même si cela a amené à une forte diminution du nombre d’observations). Trier et sélectionner les données n’est pas seulement une étape “mathématique”, les connaissances “métier” ont joué également dans le choix des variables à sélectionner pour mener à bien le projet.

Ces travaux pourraient être maintenant challengés en appliquant le modèle à une base du même type sur une autre année (plusieurs années sont disponibles sur le site de l’ADEME).

Afin d’améliorer la qualité de la prédiction des modèles, il serait possible d’analyser un dataset plus récent (après 2017). En effet, les rejets de CO2 de ce dataset ont été mesurés suivant le cycle NEDC en 2013. Il a été démontré que ce cycle n’est pas représentatif de la réalité. Les consommations et les rejets étaient nettement sous-estimés par rapport à la réalité. On peut donc penser que les conditions de réalisation de ce test engendrent un biais baissant la qualité de base des données. Ceci peut expliquer, en partie, les dispersions élevées calculées lors des régressions ainsi qu’un entraînement des modèles de classification plus difficile, baissant ainsi leurs précisions. Ce cycle de test a été remplacé en septembre 2017 par le cycle WLTP (Worldwide harmonized Light duty Test Procedure), plus proche des conditions réelles d’utilisation des véhicules.